

On Learning With Label Proportions

Felix X. Yu^{*1}, Sanjiv Kumar³, Tony Jebara², and Shih-Fu Chang¹

¹Department of Electrical Engineering, Columbia University

²Department of Computer Science, Columbia University

³Google Inc.

Abstract

We study a binary learning setting called Learning with Label Proportions (LLP), in which the training data is provided in groups, and for each group only the proportion of the positive instances are given – The task is to learn a model to predict the labels of the individual instances. LLP has broad applications in political science, marketing, healthcare, and computer vision. Though there are several works about effective algorithms for LLP, one fundamental question is left unanswered: When and why the instances labels can be learned under LLP. To answer this question, we propose a two-step analysis. We first provide a VC-type bound on the generalization error of the bag proportions. We then show that under some conditions, a good bag proportion predictor guarantees a good instance label predictor. We discuss applications of the analysis, including learning with population proportions, and justification of LLP algorithms. We also show one application of LLP, predicting income, based on real census data.

1 Introduction

A lot of information of individuals is released in the form of group label proportions. For example, after election, the proportions of votes of each demographic area are released by the government. In healthcare, the proportions of diagnosed diseases of each zip code area are available to public. Is it possible to learn a model to predict the individual labels based on only the group-level label proportions? Recent works in a machine learning setting called Learning with Label Proportions (LLP) have tried to study this problem. In LLP, the training instances are provided as bags. In a binary learning setting, for each bag, only the proportions of the positive instances are available. The task is to learn a model to predict the labels of individual instances.

It is shown that by combining the label proportions with instance-level attributes, models capable of correctly predicting instance labels can be learned [10,11,17]. As non-confidential attributes of individuals are easy to acquire, e.g., by census data, LLP not only leads to promising new applications, but also poses serious privacy concerns as releasing proportions may lead to discovery of sensitive information.

Different from the former works which focused on developing algorithms, our paper studies when and why individual labels can be learned from label proportions. Specifically, we propose a two-step analysis to answer the following questions:

- Given enough training bags, is it possible to learn a bag proportion predictor, which generalizes well to unseen bags?
- Does a “good” bag proportion predictor imply a “good” instance label predictor?

^{*}yuxinnan@ee.columbia.edu

The first question in itself is interesting as in some applications we are simply interested in getting good proportion estimates for *unseen* bags: Doctors may want to predict the rate of disease on certain geographical area, and companies may want to predict attrition rate of certain department. The second question is more crucial from privacy point of view, as the ability to learn a good instance label predictor given label proportions can be of concern.

The main result of this paper provides a formal guarantee that the answers for both of the above questions can be “yes” with high probability. In other words, We can potentially learn (and recover) the individual instance labels, by only observing label proportions. Our results hold for any algorithm.

Our work makes the following main contributions:

- We provide a VC-type bound on the generalization error of bag proportions (Section 4).
- We show that under some conditions, e.g., when instances within each bag are conditionally independent, classifiers which can achieve low bag proportion error with high probability, can also be guaranteed to achieve low instance label error. (Section 5).
- We provide discussions and experiments to demonstrate the feasibility of learning instance labels under real-world settings (Section 6, Section 7).

2 Related Works

2.1 Algorithms for LLP

In their seminal work, [10] proposed to estimate the mean of each class using the mean of each bag and the label proportions. These estimates are then used in a conditional exponential model to maximize the log likelihood. [11] proposed to use a large-margin regression method by assuming the mean instance of each bag having a soft label corresponding to the label proportion. As an extension to multiple-instance learning, [7] designed a hierarchical probabilistic model to generate consistent label proportions. Similar ideas have also been shown in [4] and [8]. A recently proposed α SVM algorithm [17] explicitly models the latent unknown instance labels together with the known group label proportions in a large-margin framework. Different from the above works, this paper provides theoretical results on when and why bag proportion and instance labels can be learned. Our results hold for all algorithms.

2.2 Privacy Attack based on Auxiliary Information

The learning with label proportion setting is related to privacy attack based on auxiliary information. Auxiliary information, a.k.a., side information or background knowledge, when combined with released statistics, can be used to compromise privacy. For example, medical records of the governor of Massachusetts were identified by linking voter registration record to “anonymized” Massachusetts Group Insurance Commission medical encounter data, which contains the DOB, sex, zip code etc [15]. As another example, sensitive user information could be recovered by combining Netflix Awards data with IMDB records [9]. In general the above attacks are achieved by linking the two information sources. While there are various definitions of data privacy, our work focuses on the *fact* that a lot of sensitive information is released in the form of label proportions. And we show that this seemingly safe way of releasing information is vulnerable, if the attacker can get hold of auxiliary information of the individuals.

2.3 Multiple Instance Learning.

A related, yet more extensively studied learning setting is Multiple Instance Learning (MIL) [5]. In MIL, the learner has access to bags, with their labels generated by the Boolean OR operator on the unobserved instance labels, i.e., a bag is positive *iff* it has at least one positive instance. The task is

to learn a binary predictor for the future *bags*. It has been shown that if all the instances are drawn iid from a single distribution, MIL is as easy as learning from iid instances with one-sided label noise [3]. In real-world applications, the instances inside each bag can have arbitrary dependencies, or even a manifold structure. The learnability and sample complexity results in the above scenarios are shown by [12,13], and [2], respectively. In this paper, we use the tools provided in [12] to analyze the generalization error of bag proportions. More importantly, we further show that under some conditions, a good bag proportion predictor implies a good instance label predictor.

3 Learning Setting And Notation

Suppose the domain of instances (attributes) is \mathcal{X} , and the domain of the instance labels is $\mathcal{Y} = \{-1, 1\}$. For simplicity, we assume that the bags are of the same size¹, and each bag is a r -tuple of instances from \mathcal{X} , $r \in \mathbb{N}$. Hence, the domain of bags is \mathcal{X}^r .

1. A bag $\tilde{x} = (x_1, \dots, x_r)$, with its corresponding instance labels $\tilde{y} = (y_1, \dots, y_r)$, are generated iid by a probability distribution D over *bags* $(\mathcal{X} \times \mathcal{Y})^r$.²

2. The learner does not have access to the instance labels. Instead it receives $(\tilde{x}, f(\tilde{y}))$, where, $f : \mathcal{Y}^r \rightarrow \mathbb{R}$, is the proportion generation function, known *a-priori*. In learning with label proportions,

$$f(\tilde{y}) = \frac{1}{r} \sum_{i=1}^r \frac{y_i + 1}{2}. \quad (1)$$

We denote the training set received by the learner as a set of m bags with proportions $S = \{(\tilde{x}_k, f(\tilde{y}_k))\}_{k=1}^m$, in which \tilde{x}_k and \tilde{y}_k are the instances and the (unobserved) labels for the k -th bag, respectively. We use $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ to denote a hypothesis class on the instances.

Definition 1. For $h \in \mathcal{H}$, define an operator to predict bag proportion based on the instances $\phi_r^f(h) : \mathcal{X}^r \rightarrow \mathbb{R}$, $\phi_r^f(h)(\tilde{x}) := f((h(x_1), \dots, h(x_r)))$, $\forall \tilde{x} \in \mathcal{X}^r$. And therefore the hypothesis class on the bags $\phi_r^f(\mathcal{H}) := \{\phi_r^f(h) | h \in \mathcal{H}\}$.

Definition 2. Given a training set S , $h \in \mathcal{H}$, denote the empirical bag proportion error with a loss function L as

$$er_S^L(h) = \frac{1}{|S|} \sum_{(\tilde{x}, f(\tilde{y})) \in S} L(\phi_r^f(h)(\tilde{x}), f(\tilde{y})). \quad (2)$$

In this paper, we consider L as the absolute loss.

Definition 3. Given $h \in \mathcal{H}$, denote the generalization error of bag proportions with a loss function L over distribution D as

$$er_D^L(h) = \mathbb{E}_{(\tilde{x}, \tilde{y}) \sim D} L(\phi_r^f(h)(\tilde{x}), f(\tilde{y})). \quad (3)$$

In section 4, we study the generalization error of bag proportions with any instance hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. In section 5, we show that, under some conditions, instance classifier h which can achieve low error of bag proportions with high probability, is guaranteed to achieve low error on instance labels with high probability.

4 Generalization Error of Bag Proportions

Before showing that good label prediction is possible for individual instances, we first show that good proportion prediction is possible for unseen bags. Our results hold for any hypotheses class.

¹Our result can be easily generalized to bags with different sizes. We will discuss this issues in Section 6.1

²Note that iid bags do not imply iid instances across all bags. Also, instances inside each bag can have arbitrary dependencies.

Note that learning the bag proportion is basically a regression problem. Therefore, for a smooth loss function L , the generalization error of bag proportions can be bounded in terms of the empirical proportion error and some complexity measure, e.g., fat shattering dimension [1], of the hypothesis class of the bag. Unfortunately, the above does not provide us insights into the LLP setting, as it does not utilize the structure of the problem. As we show later in Section 5, such structure is important in relating the error of bag proportion to the error in instance labels.

Based on the definitions in Section 3, it is intuitive that the complexity of bag proportion hypothesis class $\phi_r^f(\mathcal{H})$ should be dependent on the complexity of the instance label hypothesis class \mathcal{H} . Formally, we adapt the MIL analysis in [12], to bound the *covering number* [1] of $\phi_r^f(\mathcal{H})$, by the covering number of \mathcal{H} . As we consider the case in which \mathcal{H} is a binary hypothesis class, we further bound the covering number of \mathcal{H} based on its *VC-dimension* [16]. This leads to the following theorem on the generalization error of learning bag proportions.

Theorem 1. *For any $0 < \delta < 1$, $0 < \epsilon < 1$, $h \in \mathcal{H}$, with probability at least $1 - \delta$, $er_D^L(h) \leq er_S^L(h) + \epsilon$, if*

$$m \geq \frac{64}{\epsilon^2} (2VC(\mathcal{H}) \ln(12r/\epsilon) + \ln(4/\delta)), \quad (4)$$

in which $VC(\mathcal{H})$ is the VC dimension of the instance label hypothesis class \mathcal{H} , and m is the number of training bags.

The proof sketch is provided in the Appendix.

From the above, the generalized error of bag proportions can be bounded in terms of the empirical error if there is a sufficient number of bags. Note that the above bound is a function of bag size r , and as r grows, the problem gets harder. However the sample complexity (smallest sufficient size of m above) grows at most logarithmically with r . It means that the generalization error is mildly sensitive to r . We also note that the above theorem generalizes to the well-known case of binary supervised learning, as shown below.

Corollary 1. *When bag size $r = 1$, for any $0 < \delta < 1$, $0 < \epsilon < 1$, $h \in \mathcal{H}$, with probability at least $1 - \delta$, $er_D^L(h) \leq er_S^L(h) + \epsilon$, if*

$$m \geq \frac{64}{\epsilon^2} (2VC(\mathcal{H}) \ln(12/\epsilon) + \ln(4/\delta)). \quad (5)$$

5 Analysis of Instance Label Error

From the analysis above, we know that the generalization error of bag proportions can be bounded. The next question to answer is whether a good bag proportion predictor implies a good instance label predictor. In this section, we assume $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$ is close to 1, and discuss whether the above can guarantee low probability of misclassifying the instances, i.e., $\mathbb{P}(h(x) \neq y)$ close to 0.

Section 5.1 considers the simple case when all instances are drawn iid from a distribution. Section 5.2 considers a more general case in which instances are conditionally independent given the bag. Section 5.3 studies a scenario in which all the bags are with a fixed proportion, providing us an important failure case.

Note that the analysis in this section is based on the assumption that $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$ is close to 1. From Section 4, the above is true when we have sufficient number of training bags, and a good algorithm to achieve small empirical bag proportion error. Formally, from Theorem 1, suppose we have a learner, and some training bags, such that $\epsilon' := er_S^L(h) + \epsilon$, and

$$\mathbb{P}(er_D^L(h) \leq \epsilon') \geq 1 - \delta. \quad (6)$$

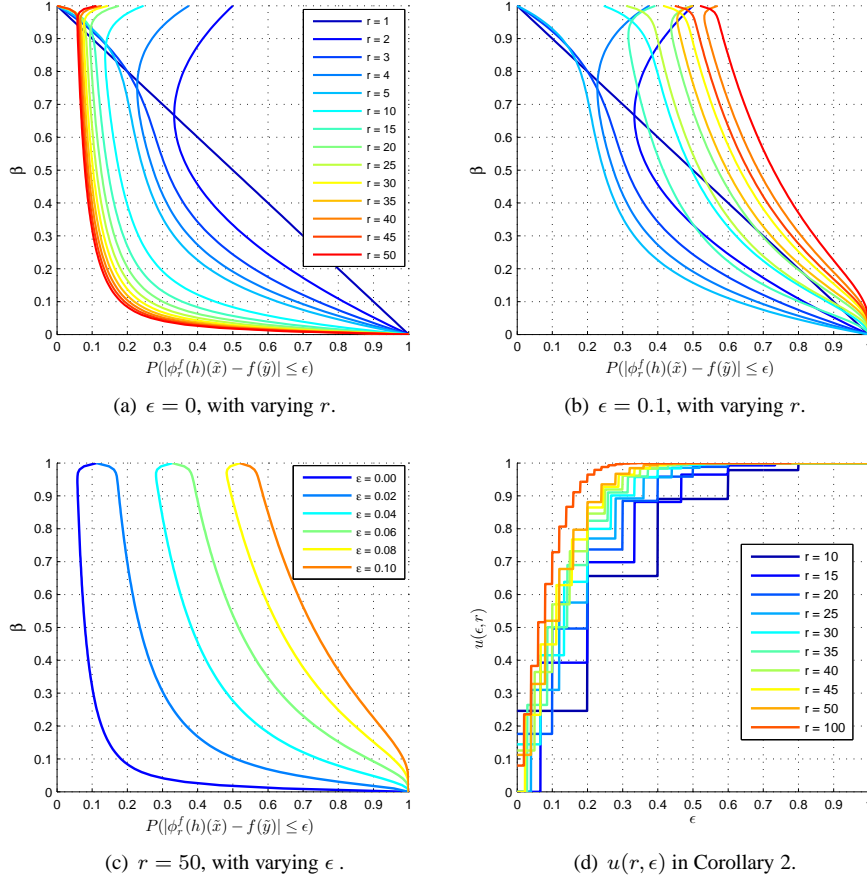


Figure 1: (a)-(c): the relationship of the instance label error β and the bag proportion error $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$, under the iid assumption of Section 5.1. (d): $u(r, \epsilon)$ in Corollary 2. Best viewed in color. (a) and (b) share the same legend.

Then $\mathbb{P}_{(\tilde{x}, \tilde{y}) \sim D}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon'')$ for some ϵ'' can also be bounded, for example, with the Markov's Inequality. Suppose $\mathbb{P}(er_D^L(h) \leq \epsilon') \geq 1 - \delta$. For any $t > 0$, define $\epsilon'' = t\epsilon'$. With probability at least $1 - \delta$,

$$\mathbb{P}_{(\tilde{x}, \tilde{y}) \sim D}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon'') \geq 1 - \frac{1}{t}. \quad (7)$$

5.1 All Instances Are Drawn IID

We first consider a very simple scenario, in which all the instances are drawn iid from a single distribution. To simplify the analysis, we assume the prior of the instances can be matched by the hypothesis, i.e.,

$$\mathbb{P}(h(x) = 1) = \mathbb{P}(y = 1) = \alpha \quad (8)$$

The above assumption is not too restrictive given sufficient number of training bags, and an algorithm which can achieve low bag proportion error. Denote the instance label error as $\beta := \mathbb{P}(h(x) \neq y)$. The following theorem relates β to $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$.

Theorem 2. When all instances are drawn iid, the relation of instance label error β and $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$ is

$$\begin{aligned} & \mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon) \\ &= a^r \sum_{i=0}^r \binom{r}{i} b^i \left(\mathcal{F}(i + \lfloor \epsilon r \rfloor; r - i, b) - \mathcal{F}(i - \lfloor \epsilon r \rfloor - 1; r - i, b) \right), \end{aligned} \quad (9)$$

where $\lfloor \cdot \rfloor$ is the floor operator, $a = (2 - \beta)/2$, $b = \beta/(2 - \beta)$, $0 < \beta < 1$, $0 < \epsilon < 1$ and \mathcal{F} is the CDF of binomial distribution.

Proof. Based on the assumption in (8) and the definition of β ,

$$\begin{aligned} \mathbb{P}(h(x) = 1 \wedge y = 1) &= \alpha - \beta/2, \\ \mathbb{P}(h(x) = 1 \wedge y = 0) &= \beta/2, \\ \mathbb{P}(h(x) = 0 \wedge y = 1) &= \beta/2, \\ \mathbb{P}(h(x) = 0 \wedge y = 0) &= 1 - \alpha - \beta/2. \end{aligned}$$

Therefore we can explicitly write down $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$ as a function of β . The final result is obtained by collecting terms and utilizing the fact that for $m, n \in \mathbb{N}$, $m \leq n$, $\lambda \geq 0$, $\sum_{k=0}^m \binom{n}{k} \lambda^k = (1 + \lambda)^n \mathcal{F}\left(m; n, \frac{\lambda}{1 + \lambda}\right)$. \square

Note that the relationship in Theorem 2 is independent of α . We will later show in Section 5.2 that this important fact makes it possible to generalize the theorem to a more relaxed case requiring only instances to be conditionally independent given the bag.

We have expressed $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$ as a function of β . However, what we really want to have is to control β by $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$. While the general form is hard to derive analytically, we draw the “inverse” function in Figure 1, with the following observations:

1. When $r = 1$, the misclassification error is simply $\beta = 1 - \mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$ (i.e., the case of binary supervised learning).
2. The function in Theorem 2 is monotonically decreasing when r is odd, and the function is not monotonic, when r is even. This can be explained by the order of the polynomial, as an odd r will lead to a polynomial with odd order, and an even r otherwise. For example, considering the simple case when $\epsilon = 0$. $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon) = -\beta + 1$, for $r = 1$, and $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon) = -\frac{3}{2}\beta^2 - 2\beta + 1$, for $r = 2$.
3. The function in Theorem 2 cannot be inverted in general. However, for all r , it can be inverted when $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$ is sufficiently close to 1. For example, in Figure 1 (a), when $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon) \in (0.5, 1]$, larger $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$ leads to smaller β .
4. From Figure 1 (b) (c), as we increase ϵ , all the curves shift to right, and curves with larger r is more sensitive to ϵ . Note that curves for $r \in \{1, 2, 3, 4, 5, 10\}$ in (b) is the same as the ones in (a), since the proportion can only be discrete values in $\{0, 1/r, 2/r, \dots, 1\}$.

The result below justifies the above observations.

Corollary 2. β is a monotonically decreasing function of $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$, if

$$\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon) \in (u(r, \epsilon), 1], \quad (10)$$

where,

$$u(r, \epsilon) = \left(\frac{1}{2}\right)^r \sum_{i=\lceil (r - \lfloor \epsilon r \rfloor)/2 \rceil}^{\lfloor (r + \lfloor \epsilon r \rfloor)/2 \rfloor} \binom{r}{i}. \quad (11)$$

Specifically, $u(r, 0) = 0$, when r is odd. And $u(r, 0) = \binom{r}{r/2} \left(\frac{1}{2}\right)^r$, when r is even.

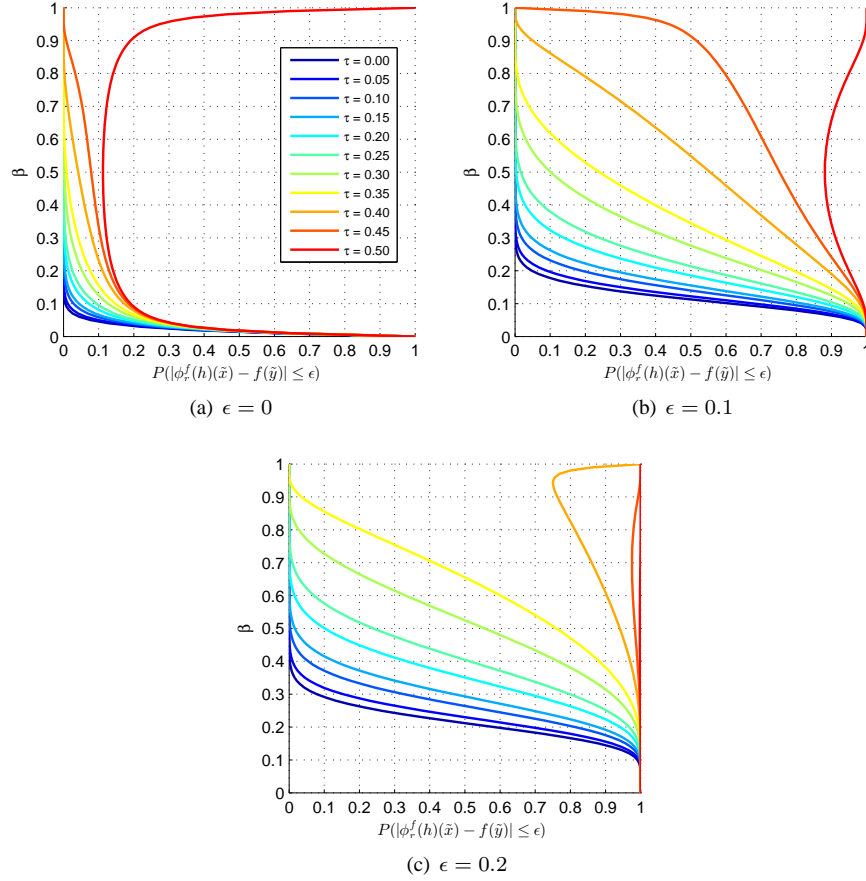


Figure 2: The relationship of β and $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$ with varying bag proportion $\tau = f(\tilde{y})$, under the assumptions of Section 5.3. Best viewed in color. The figure is generated based on $r = 50$. Note that the relationship in Theorem 3 is symmetric to $\tau = 0.5$, so we only consider τ from 0 to 0.5.

The curve of $u(r, \epsilon)$ is shown in Figure 1 (d). A larger ϵ leads to a larger $u(r, \epsilon)$, and $u(r, \epsilon)$ of larger r is more sensitive to ϵ .

In conclusion, under the iid instance assumption, an instance hypothesis, which can achieve sufficiently good bag proportion prediction, guarantees low instance label error, and hence can lead to privacy concerns.

5.2 Instances Inside Each Bag Are Conditionally Independent

The analysis in the previous section was based on a strong assumption that all instances were generated iid. In this section, we relax the assumption and generalize the results to the case where the instances are conditionally independent given the bag. A lot of real-world applications follow this assumption. For example, in modeling voting behavior, each bag is generated by randomly sampling a number of individuals from certain location. It is reasonable to assume individuals are iid given location. Similarly in healthcare, certain disease can be seen as highly geographically dependent.

Suppose the bags are generated from a distribution over bags \mathcal{D} . \mathcal{D} is a mixture of multiple components D_1, \dots, D_n , where each D_i is a distribution over bags. To make this easier to understand, we

can consider drawing a bag from \mathcal{D} as firstly picking a distribution D_i with some fixed probability θ_i , and then generating a bag from D_i .

We assume for $D_i, i = 1, \dots, n$, there exists an instance distribution D'_i for each D_i , such that generating a bag from D_i is by drawing r iid instances from D'_i .

If we have sufficient number of training bags drawn from \mathcal{D} , and a good learner of label proportions, it is fair to assume that the prior of each instance distribution can be matched, i.e., $\mathbb{P}_{(x,y) \sim D'_i}(h(x) = 1) = \mathbb{P}_{(x,y) \sim D'_i}(y = 1) = \alpha_i, \forall i = 1, \dots, n$, in which α_i is the prior of D'_i .

Define $\beta_i = \mathbb{P}_{(x,y) \sim D'_i}(h(x) \neq y)$. Then Theorem 2 is true for β_i and $\mathbb{P}_{(\tilde{x}, \tilde{y}) \sim D_i}(|\phi_r^f(h)(\tilde{x}) - f(\tilde{y})| \leq \epsilon)$, for all $i = 1, \dots, n$. We further note that the relationship is independent of α_i . Therefore Theorem 2 and all the analysis in Section 5.1 can be directly applied to $(\tilde{x}, \tilde{y}) \sim \mathcal{D}$.

5.3 All Bags with A Fixed Proportion

We consider a very different setting, in which all the bags have a fixed proportion, denoted as τ . Intuitively, if a bag proportion predictor works perfectly for all possible bags with proportion exactly 0.5, the predicted label can either be 100% correct, or 100% wrong. This provides us a failure case for achieving good instance label prediction. On the other hand, if we have a perfect proportion prediction on all possible bags of proportion 0 and 100%, we are certain that the instance labels are all correct. This section helps to understand the above intuition.

Suppose that a bag is formed by first drawing $r^+ = r\tau$ iid positive instances from a positive distribution D_+ , and then drawing $r^- = r(1 - \tau)$ iid negative instances from a negative distribution D_- . In other words, the distribution over the bags in such case is $D_+^{r^+} \times D_-^{r^-}$.

For simplicity, we assume that the conditional probability of making wrong instance label prediction is balanced, i.e.,

$$\mathbb{P}(h(x) = 0|y = 1) = \mathbb{P}(h(x) = 1|y = 0) = \beta \quad (12)$$

Theorem 3. *When all bags are drawn by the above method with fixed proportions:*

$$\begin{aligned} & \mathbb{P}(|\phi_r^f(h)(\tilde{x}) - \tau| \leq \epsilon) \\ &= \beta^{r^+} (1 - \beta)^{r^-} \sum_{(i,j) \in \mathcal{K}} \binom{r^+}{i} \binom{r^-}{j} \left(\frac{1 - \beta}{\beta}\right)^{i-j}, \end{aligned} \quad (13)$$

where $\mathcal{K} = \{(i, j) | i + j \in [r^+ - \lfloor r\epsilon \rfloor, r^+ + \lfloor r\epsilon \rfloor], i \geq 0, j \geq 0, r^+ \geq i, r^- \geq j\}, \beta \in [0, 1]$.

The proof of the above is similar to that of Theorem 2. Note that the above is symmetric to $\tau = 0.5$. So in Figure 2, we only draw τ from 0 to 0.5.

Same as in Section 5.1, we are actually interested in expressing β as a function of $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - \tau| \leq \epsilon)$. We can see from Figure 2 that except τ close to 0.5, a larger $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - \tau| \leq \epsilon)$ leads to a smaller β . In fact, when τ is close to 0.5, $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - \tau| \leq \epsilon) \rightarrow 1$ will either lead to $\beta \rightarrow 0$ or $\beta \rightarrow 1$ as there is no way for the learner to determine which “side” is positive. In addition, the curve becomes sharper when τ is closer to 0 or 1. We have the following result to justify the above observation.

Corollary 3. *Under the above assumption, β is a monotonically decreasing function of $\mathbb{P}(|\phi_r^f(h)(\tilde{x}) - \tau| \leq \epsilon)$, if*

$$\tau < (1 - \epsilon)/2 \quad \text{or} \quad \tau > (1 + \epsilon)/2. \quad (14)$$

The result provided in this section is interesting in two ways. On the one hand, it provides us a failure case, in which even perfect bag proportion prediction does not guarantee low error on the instance labels. On the other hand, it provides guidance when releasing sensitive information. For example, it might be safer for the curator to release bag proportions closer to 0.5, compared to proportions closer to 0 or 1.

6 Discussions

Section 6.2 discusses the connection of existing LLP algorithms to our analysis. Section 6.3 discusses an application called learning with “population” proportions, when the actual bag proportions are not observed.

6.1 Bags with Different Sizes

Our result can be easily generalized to bags with different sizes. Theorem 1 also holds when the bag sizes are different by simply replacing r with the average bag size. This is based on the fact that Lemma 1 holds with average bag size. Similarly, in Section 5, we have shown that for *any* bag size, low bag proportion error can guarantee low instance error.

6.2 LLP Algorithms

Among the existing algorithms, our analysis is well aligned with algorithms which learn an instance label classifier by minimizing the empirical bag proportion error, with proper regularization. Such algorithms are intuitive for LLP, as we cannot directly minimize the empirical label error with only label proportions. One algorithm in this category is \propto SVM [17], which has been shown to give good performance. The objective of \propto SVM is to find a large-margin classifier consistent with the label proportions. The analysis of Section 4 can be generalized to justify \propto SVM based on margin analysis, e.g., [18].

Note that there are algorithms, e.g., Meanmap [10] and Inverse Calibration [11], whose objectives are not to directly minimize the empirical bag proportion error. The analysis can still be applied by first computing the empirical bag proportion error based on the instance label classifier $h \in \mathcal{Y}^{\mathcal{X}}$ learned by the algorithm.

6.3 Learning with Population Proportions

In real-world settings, sometimes a proportion is given based on all instances in one location, but we are only given a small sample of these instances per location at training time. Note that location can be substituted by other attributes, such as ethnic groups. Consider the scenario of modeling voting behavior based on government released statistics, the government releases the population proportion (e.g. 62.6% in New York voted for Obama in 2012 election) of each location, and we only have a subset of randomly sampled instances for each location. Can LLP be applied to correctly predict labels of the individuals? By treating each location as a bag, algorithms, e.g. \propto SVM, can only minimize the proportion error in terms of the population proportions, because the actual proportions of the selected subsets are not available. Similar to Section 5.2, one population proportion can be seen as a “true proportion” for a distribution over bags, in this case, a location. The training bags are sampled from a mixture of locations.

Suppose a bag is formed by randomly sampling r instances $\{(x_i, y_i)\}_{i=1}^r$ from a location with a true proportion p^* . We can apply the Chernoff bound to the random variables $(y_i + 1)/2, i = 1 \cdots r$: $\mathbb{P}(|f(\tilde{y}) - p^*| \geq \epsilon) \leq 2e^{-2r\epsilon^2}$. In other words, when $r \geq \ln(2/\delta)/(2\epsilon^2)$, with probability at least $1 - \delta$, $|f(\tilde{y}) - p^*| \leq \epsilon$.

We can connect the above with Theorem 1, by treating the bags as iid, and assuming that there is a gap between the actual proportion (unobserved) and the population proportion characterized above. With triangle inequality, we then have the following result.

Corollary 4. *Suppose $er_S^L(f)^*$ is the empirical error with the true proportion. For any $0 <$*

Grouping Attribute	Native Country	Occupation	Education	Race	Oracle
# Bags	41	15	15	5	-
Test Prop. Err (%)	11.94 \pm 0.93	9.84 \pm 0.73	16.65 \pm 0.39	24.34 \pm 0.62	-
Test Instance Err (%)	18.75 \pm 0.25	18.19 \pm 0.16	19.61 \pm 0.10	24.02 \pm 0.15	15.07 \pm 0.07

Table 1: Error on predicted income on adult dataset. We randomly select 80% of the individuals for training, and 20% for testing. The parameters are tuned based on cross validation on the training bags. The result is based on 5 experiments. We report mean with standard deviation. We report an “oracle result”, which is based on linear SVM with access to all the labels of the training instances. This can be seen as a performance upper bound.

$\epsilon_1, \epsilon_2, \delta_1, \delta_2 < 1$, with probability at least $(1 - \delta_1)(1 - \delta_2)$, $er_D^L(f) - er_S^L(f)^* \leq \epsilon_1 + \epsilon_2$, if

$$m \geq \frac{64}{\epsilon_1^2} (2VC(\mathcal{H}) \ln(12r/\epsilon_1^2) + \ln(4/\delta_1)), \quad (15)$$

$$r \geq \ln(2/\delta_2)/(2\epsilon_2^2). \quad (16)$$

7 Experiments

Until now, we have provided a formal analysis that under some conditions, labels of individual instances can indeed be learned or recovered.

In this section, we conduct an experiment with real-world data to demonstrate the feasibility of getting a good instance classifier under the LLP setting.

The experiment is based on \propto SVM³ with linear kernel, and absolute loss.

We conduct experiments to predict the household income based on census. The dataset, “adult”, was extracted from 1994 census data. It contains 32,561 instances, each with 123 binary attributes, including education, marital status, sex etc. Each individual has a binary label indicating whether “income > 50k”.

We consider “income > 50k” as sensitive information, for which we only know its proportions on different bags. The bags are formed based on a “grouping attribute”. For example, if the grouping attribute is “native country”, we group the individuals into 41 bags corresponding to the 41 native countries. Other grouping attributes we considered are “occupation”, “education”, and “race”.

In the real-world applications, “income > 50k” could be substituted by voting behavior, attrition, diagnosed disease etc., and the government or some other organizations may release the proportions based on different grouping attributes including location, age, mortgage or the ones we are using. The attributes of individuals could be acquired from census data or surveys.

In the experiment, we use 80% of the data for training and 20% for testing. For the training data, we assume the proportion of each bag to be known. Table 1 shows the performance of LLP in terms of the test bag proportion error, and test instance label error. We find that the performance of LLP is not much worse compared to oracle (and certainly much better than random guess, or guessing with prior). Note that in the real-world case, we may only know the population proportion (Sectino 6.3), and the performance of LLP may drop depending on the closeness of the actual proportion to the population proportion. We also observed that the predictor learned by \propto SVM is quite intuitive. For example, people with higher education or higher working hours are more likely to have higher income.

³<https://github.com/felixyu/pSVM>

8 Conclusion

This paper proposed a novel two-step analysis to answer the question when and why individual labels can be learned in the LLP setting. Specifically, we showed how parameters such as bag size and bag proportion affect the bag proportion error and instance label error. The generalization error of bag proportions is only mildly sensitive to the size of the bags. Under some general conditions, e.g. instances inside each bag are conditionally independent, a good bag proportion predictor guarantees a good instance label predictor. Experiments with real-world data verified the theoretical analysis.

There are several directions for future works. Some alternative tools, e.g. sample complexity results of learning $\{0, \dots, n\}$ -valued functions [6], can be used for analyzing the generalization error of bag proportions. Data dependent measure e.g., Rademacher complexity, may lead to tighter bound for practical use. We also believe that the assumptions of Section 5 can be considerably relaxed to match more complex real-world applications.

9 Appendix

9.1 Proof Sketch of Theorem 1

One important tool used in the proof is the lemma below bounding the covering number of bag proportion hypothesis class $\phi_r^f(\mathcal{H})$ by the covering number of the instance hypothesis class \mathcal{H} .

Lemma 1. [13] *Let $r \in \mathbb{N}$ and suppose $f : \mathbb{R}^r \rightarrow \mathbb{R}$ is α -Lipschitz w.r.t. the infinity norm, for some $\alpha_f > 0$. Let $\gamma > 0$, $p \in [1, \infty]$, and $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$. For any $m \geq 0$,*

$$\mathcal{N}_p(\gamma, \phi_r^f(\mathcal{H}), m) \leq \mathcal{N}_p\left(\frac{\gamma}{\alpha_f r^{1/p}}, \mathcal{H}, rm\right). \quad (17)$$

Covering number [1] can be seen as a complexity measure on real-valued hypothesis class. The larger the covering number, the larger the complexity. Another lemma we use is the uniform convergence for real function class.

Lemma 2. [1]. *Let $\hat{\mathcal{Y}}, \mathcal{Y} \subseteq \mathbb{R}$, $\mathcal{G} \subseteq \hat{\mathcal{Y}}^{\mathcal{X}}$, and $L : \hat{\mathcal{Y}} \times \mathcal{Y} \rightarrow [0, 1]$, such that L is Lipschitz in its first argument with Lipschitz constant $\alpha_L > 0$. Let D be any distribution on $\mathcal{X} \times \mathcal{Y}$. Then for any $0 < \epsilon < 1$ and $g \in \mathcal{G}$:*

$$\mathbb{P}_{S \sim D^m} \left(\sup_{g \in \mathcal{G}} |er_D^L(g) - er_S^L(g)| \geq \epsilon \right) \quad (18)$$

$$\leq 4\mathcal{N}_1(\epsilon/(8\alpha_L), \mathcal{G}, 2m)e^{-m\epsilon^2/32}, \quad (19)$$

in which $er_S^L(g) = \frac{1}{|S|} \sum_{x \in S} L(g(x), y)$, $er_D^L(g) = \mathbb{E}_{x \sim D} L(g(x), y)$.

Combining the fact $d_2(x, x') < d_\infty(x, x')$, $\forall x, x'$, where d_2 and d_∞ represent the ℓ_2 and ℓ_∞ distance, respectively, and the definition of covering number: $\mathcal{N}_1(\epsilon, W, m) \leq \mathcal{N}_\infty(\epsilon, W, m)$. Applying Lemma 1:

$$4\mathcal{N}_1(\epsilon/(8\alpha_L), \bar{\mathcal{H}}, 2m)e^{-m\epsilon^2/32} \quad (20)$$

$$\leq 4\mathcal{N}_\infty(\epsilon/(8\alpha_L), \bar{\mathcal{H}}, 2m)e^{-m\epsilon^2/32} \quad (21)$$

$$\leq 4\mathcal{N}_\infty(\epsilon/(8\alpha_L\alpha_f), \mathcal{H}, 2rm)e^{-m\epsilon^2/32}. \quad (22)$$

Also, the proportion generation function f in (1) is 1-Lipschitz with the infinity norm. And the loss functions L we are considering is 1-Lipschitz.

Based on the definition of covering number, for $\mathcal{H} \subseteq \{-1, 1\}^{\mathcal{X}}$, for any $\epsilon < 2$, $\mathcal{N}(\epsilon, \mathcal{H}|_{x_1^m}, d_\infty) = |\mathcal{H}|_{x_1^m}|$. Thus,

$$\mathcal{N}_\infty(\epsilon, \mathcal{H}, m) = \Pi_{\mathcal{H}}(m). \quad (23)$$

Refer to [1] for the definition of *restriction* $\mathcal{H}|_{x_1^m}$ and *growth function* $\Pi_{\mathcal{H}}(m)$. In addition, we have the following lemma to bound the growth function by VC dimension of the hypothesis class.

Lemma 3. [14] Let $\mathcal{G} \subseteq \{-1, 1\}^{\mathcal{X}}$ with $VC(\mathcal{G}) = d \leq \infty$. For all $m \geq d$,

$$\Pi_{\mathcal{G}}(m) \leq \sum_{i=0}^d \binom{m}{i} \leq \left(\frac{em}{d}\right)^d. \quad (24)$$

Let $d = VC(\mathcal{H})$. By combining the above facts, and $0 < \epsilon < 1$, (22) leads to

$$4\Pi_{\mathcal{H}}(2rm)e^{-m\epsilon^2/32} \leq 4\left(\frac{2erm}{d}\right)^d e^{-m\epsilon^2/32}. \quad (25)$$

Therefore, with probability at least $1 - \delta$,

$$\begin{aligned} \text{er}_D^L(f) &\leq \text{er}_S^L(f) + \left(\frac{32}{m}(d \ln(2erm/d) + \ln(4/\delta))\right)^{1/2} \\ &\Leftarrow m \leq \frac{32}{\epsilon^2}(d \ln m + d \ln(2er/d) + \ln(4/\delta)). \end{aligned} \quad (26)$$

Since $\ln x \leq ax - \ln a - 1$ for all $a, x > 0$, we have

$$\Leftarrow \frac{32d}{\epsilon^2} \ln m \leq \frac{32d}{\epsilon^2} \left(\frac{\epsilon^2}{64d} m + \ln \left(\frac{64d}{\epsilon^2} \right) \right) \quad (27)$$

$$\leq \frac{m}{2} + \frac{32d}{\epsilon^2} \ln \left(\frac{64d}{e\epsilon^2} \right). \quad (28)$$

$$\Leftarrow m \geq \frac{m}{2} + \frac{32}{\epsilon^2}(d \ln(128r/\epsilon^2) + \ln(4/\delta)). \quad (29)$$

$$\Leftarrow m \geq \frac{64}{\epsilon^2}(2d \ln(12r/\epsilon) + \ln(4/\delta)). \quad (30)$$

References

- [1] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.
- [2] Boris Babenko, Nakul Verma, Piotr Dollár, and Serge J Belongie. Multiple instance learning with manifold bags. In *Proceedings of the 28th International Conference on Machine Learning*, pages 81–88, 2011.
- [3] Avrim Blum and Adam Kalai. A note on learning from multiple-instance examples. *Machine Learning*, 30(1):23–29, 1998.
- [4] B.C. Chen, L. Chen, R. Ramakrishnan, and D.R. Musicant. Learning from aggregate views. In *Proceedings of the 22nd International Conference on Data Engineering*, pages 3–3. IEEE, 2006.
- [5] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997.

- [6] David Haussler and Philip M Long. A generalization of Sauer’s Lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.
- [7] Hendrik Kuck and Nando de Freitas. Learning about individuals from group statistics. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 332–339. AUAI Press, 2005.
- [8] D.R. Musicant, J.M. Christensen, and J.F. Olson. Supervised learning by training on aggregate outputs. In *Proceedings of the 7th International Conference on Data Mining*, pages 252–261. IEEE, 2007.
- [9] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125. IEEE, 2008.
- [10] N. Quadrianto, A.J. Smola, T.S. Caetano, and Q.V. Le. Estimating labels from label proportions. *The Journal of Machine Learning Research*, 10:2349–2374, 2009.
- [11] S. Rüeping. SVM classifier estimation from group probabilities. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.
- [12] Sivan Sabato. *Partial Information and Distribution-Dependence in Supervised Learning Models*. PhD thesis, School of Computer Science and Engineering, Hebrew University of Jerusalem, 2012.
- [13] Sivan Sabato and Naftali Tishby. Multi-instance learning with any hypothesis class. *Journal of Machine Learning Research*, 13:2999–3039, 2012.
- [14] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.
- [15] Latanya Sweeney. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997.
- [16] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.
- [17] F. X. Yu, D. Liu, Sanjiv K., T. Jebara, and S.-F. Chang. ∞ SVM for learning with label proportions. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.
- [18] Tong Zhang. Covering number bounds of certain regularized linear function classes. *The Journal of Machine Learning Research*, 2:527–550, 2002.